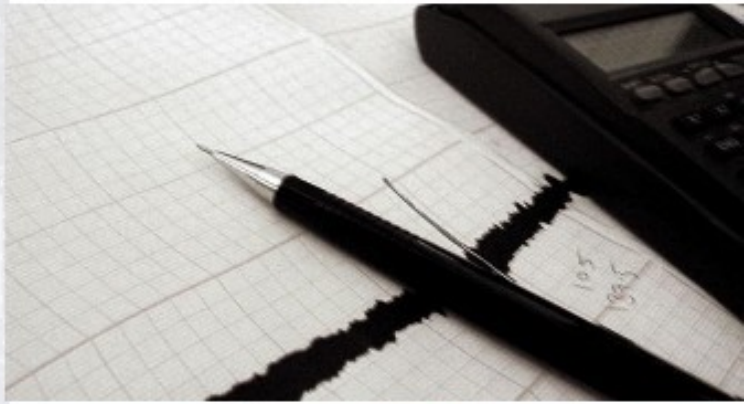
zalando
| RESEARCH |

FASHION DNA



Christian Bracher

KDD 2016 Fashion Workshop
San Francisco, August 2016

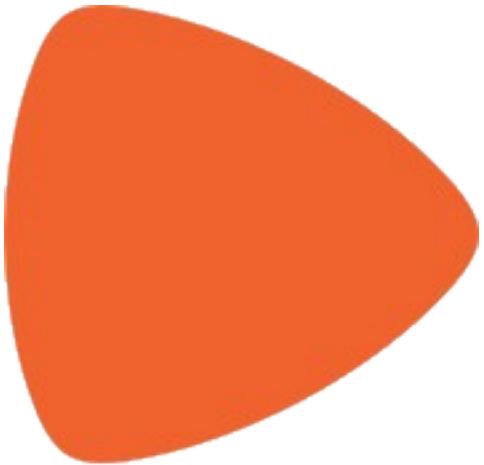
Company Information

- Europe's leading online fashion platform
- Operating in 15 countries

- 920 million € sales in Q2/2016
- 11,000 employees from 100+ countries

- 150,000 fashion items offered at any time
- 19 million active customers

- 1,300 employees at Zalando Technology
- 75 data scientists and growing



Interested? **Visit:** https://tech.zalando.de/ and https://jobs.zalando.de/en/

Integrating Sales and Content Data

We accumulated a treasure trove of data:

- Article information for ~10^6 articles (SKUs)
- Details for ~10^8 completed sales events
- Involving ~$4 \cdot 10^7$ customers

Goals of merging sales data and article information:

- Article maps: Integrating user and expert views of fashion
- Improved personalized recommendations

The Fashion DNA Approach: Basic Idea

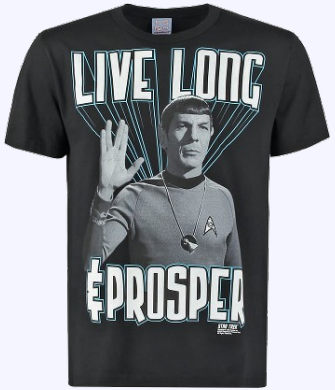
Represent items and customers as vectors in dual spaces:

Fashion DNA vector $\mathbf{f}_v$

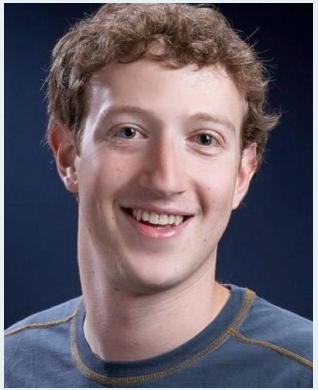
Customer Style vector $\mathbf{s}_k$

The Fashion DNA Approach: Basic Idea

Cosine similarity $\mathbf{f}_\mu \cdot \mathbf{f}_\nu$



Customer–Item affinity $\mathbf{f}_\nu \cdot \mathbf{s}_k$



Cosine similarity $\mathbf{s}_j \cdot \mathbf{s}_k$

Express similarity and affinity by inner products

The Fashion DNA Approach: Basic Idea

Customer–item affinity should underline probability of sale:

$$p_{\nu k} = \sigma(\mathbf{f}_\nu \cdot \mathbf{s}_k + \beta_k)$$

- σ(...) : logistic function
- β_k : customer bias (buying propensity)

- Compare to "ground truth": Binary matrix Π of actual sales
- Goal: Adjust $\mathbf{f}_\nu$, $\mathbf{s}_k$, β_k to minimize average logistic loss $\bar{L}$:

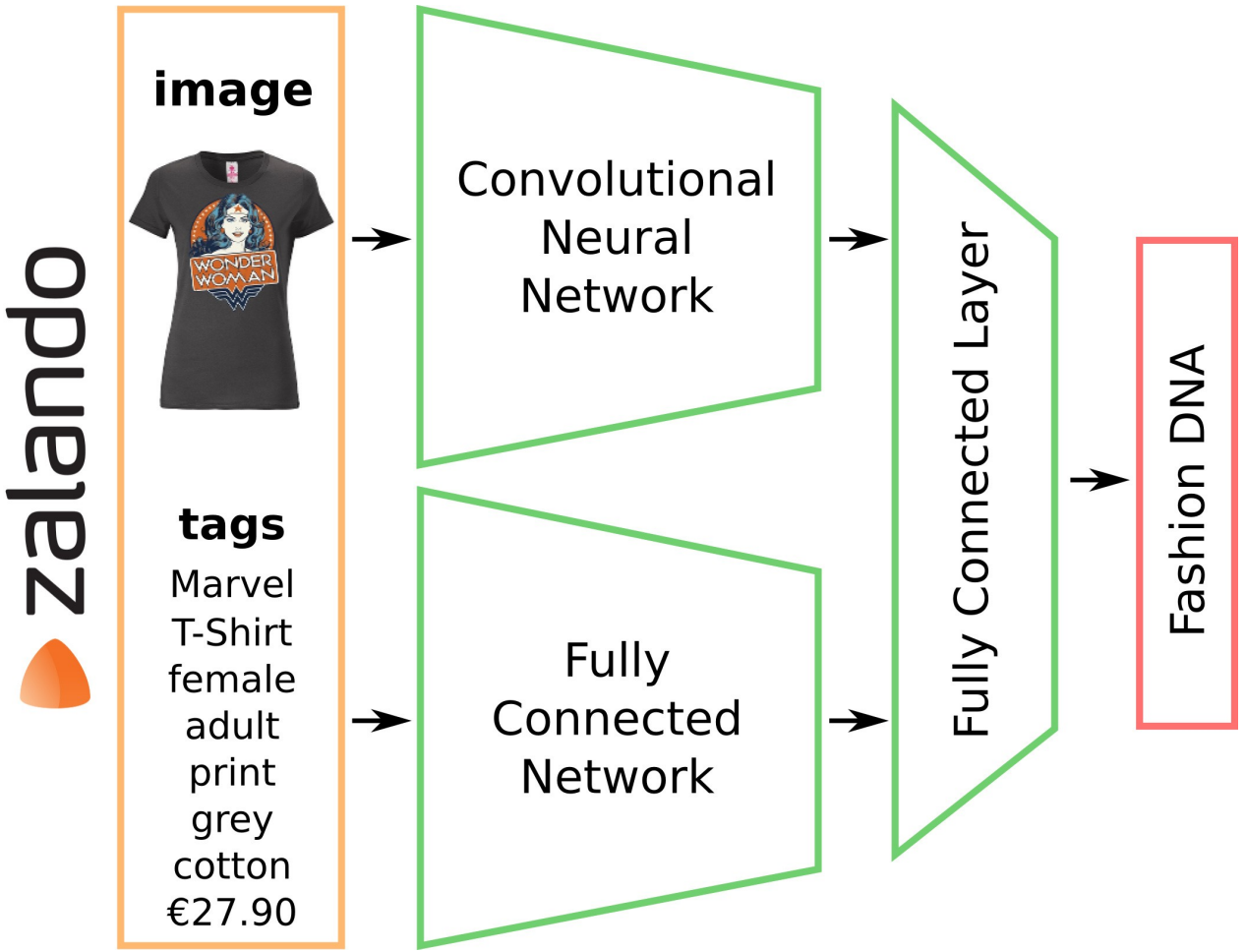
$$\bar{L} \propto \sum_\nu \sum_k [\pi_{\nu k} \log p_{\nu k} + (1-\pi_{\nu k}) \log(1-p_{\nu k})]$$

C. Johnson, *Logistic matrix factorization for implicit feedback data*, in *NIPS Workshop on Distributed Matrix Computations*, 2014.

Boosting Fashion DNA

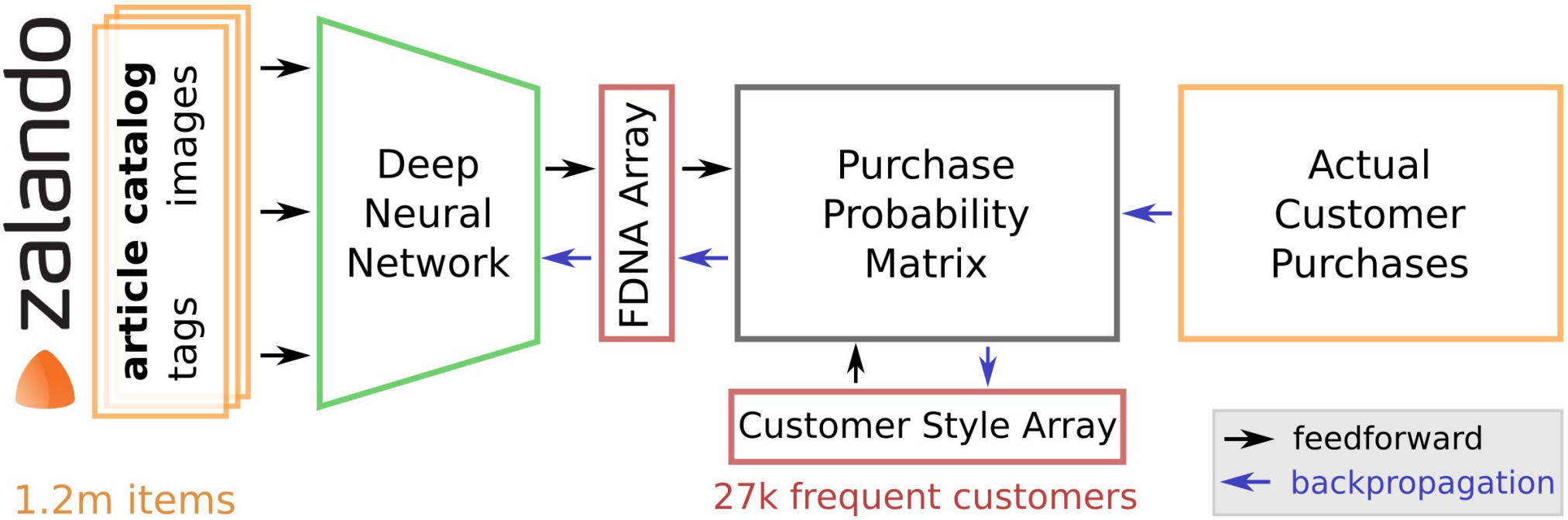
- So far, a sales-based logistic matrix factorization model
- Not really adequate in the fashion context:
  - Rapid turnover of inventory: ~10^3 new articles/day
  - Ignores curated information about fashion items
  - Sales events are sparse: Cold-start problem
- Model requirements:
  - Assign Fashion DNA to articles in the absence of sales
  - Assign style vectors to active customers on-demand

From Curated Content to Fashion DNA



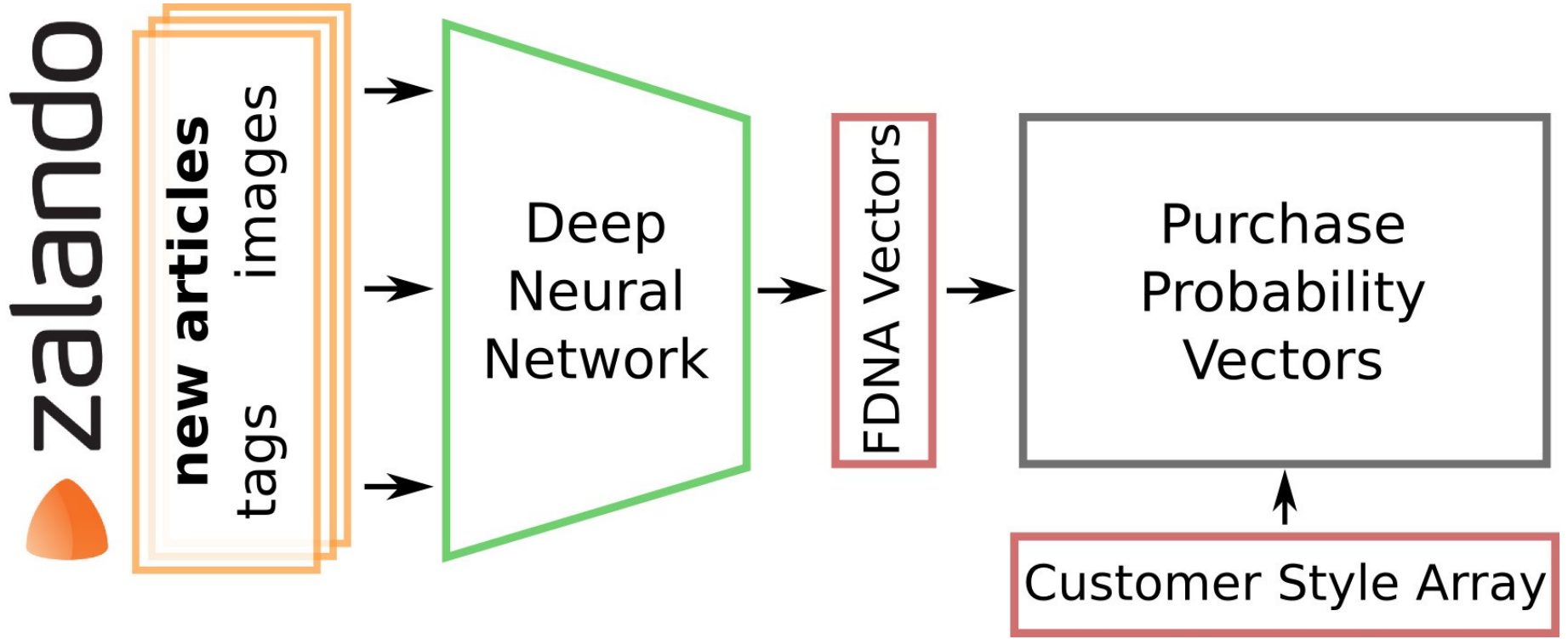
- "Clamp" curated item data to sales information via deep learning

Training the Model on Sales Data



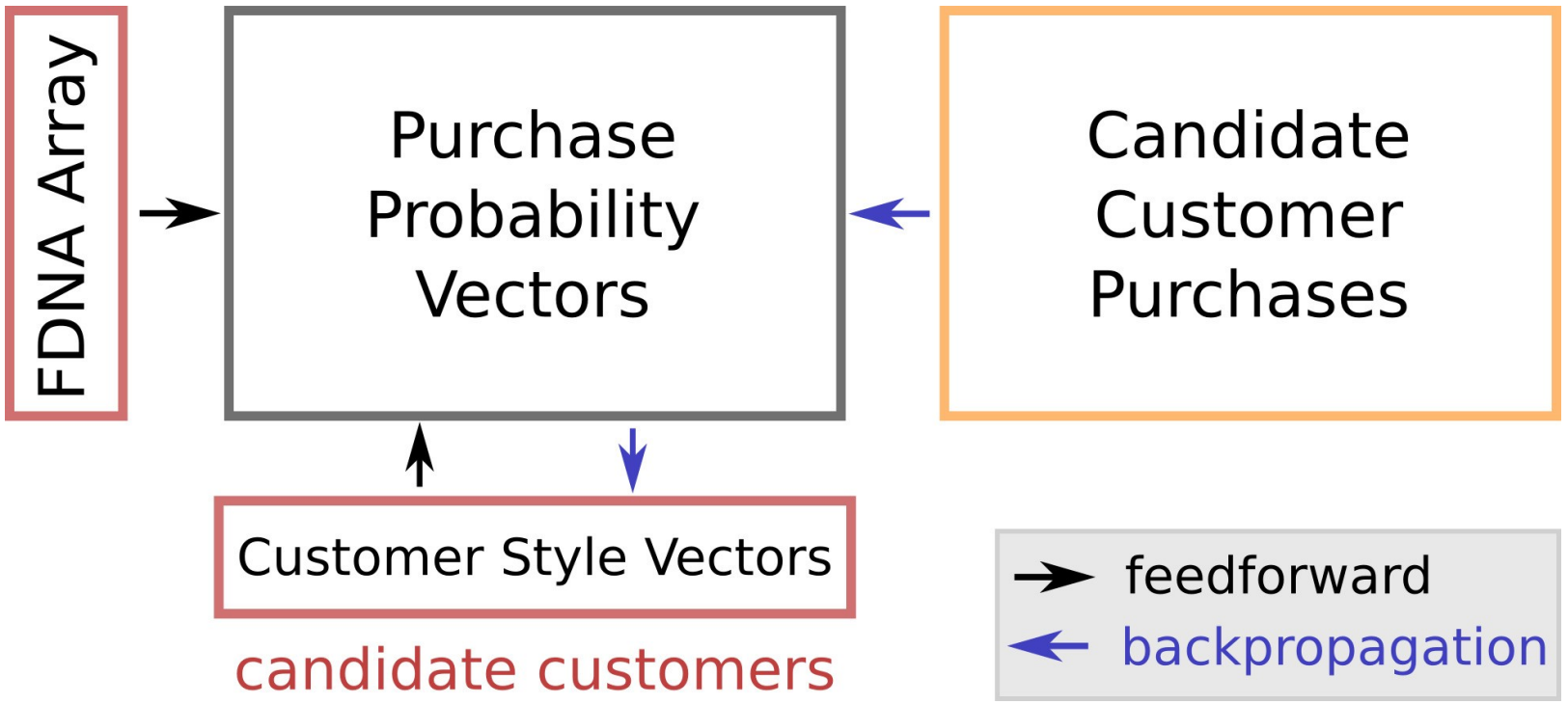
- Train network weights and style vectors/bias for frequent customers

Recommending New Fashion Items



- Pass curated information through trained network

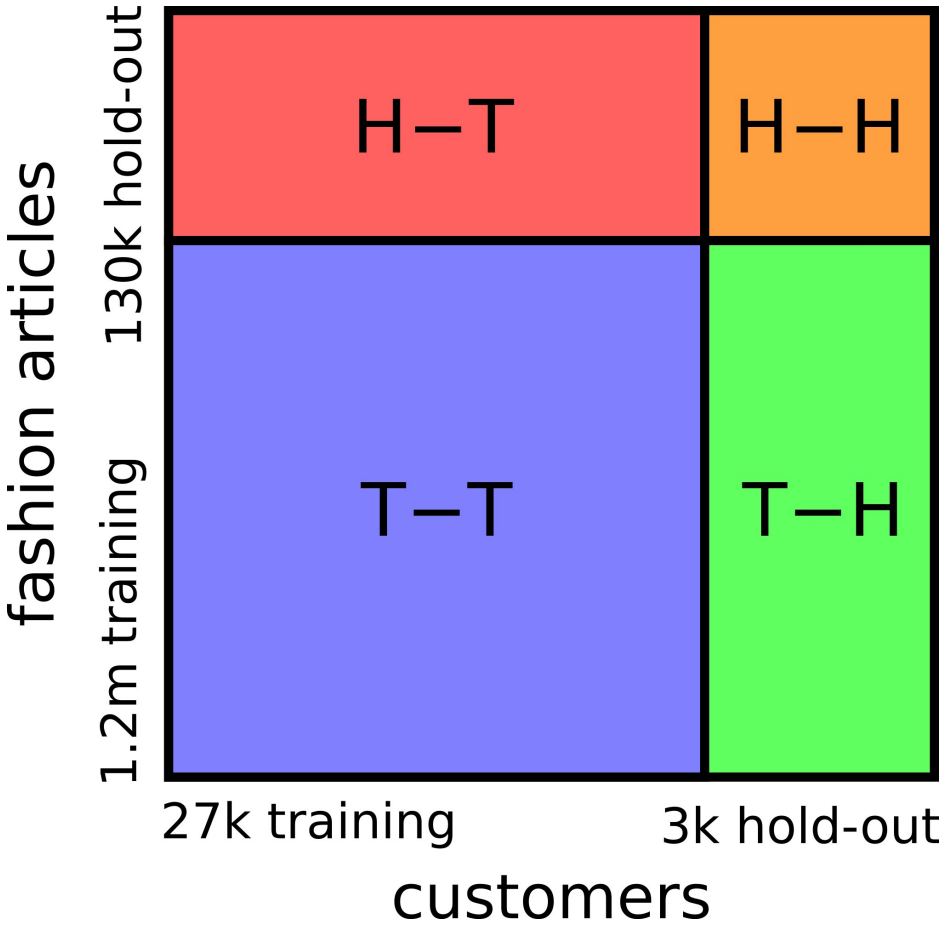
Assigning Styles to Hold-Out Customers



- Train customer style vectors/biases for fixed Fashion DNA array

Evaluating Recommendations

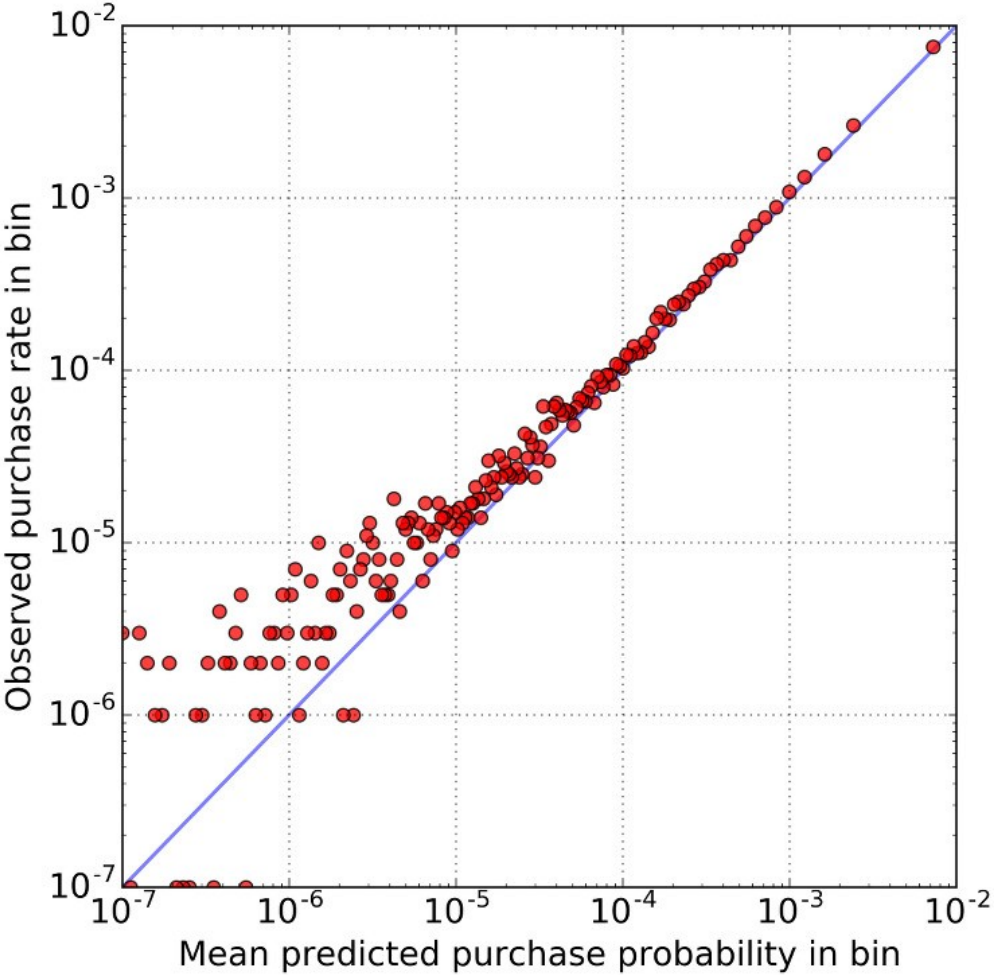
- 1.3m fashion articles
- 30k frequent customers
- Both split randomly 90:10



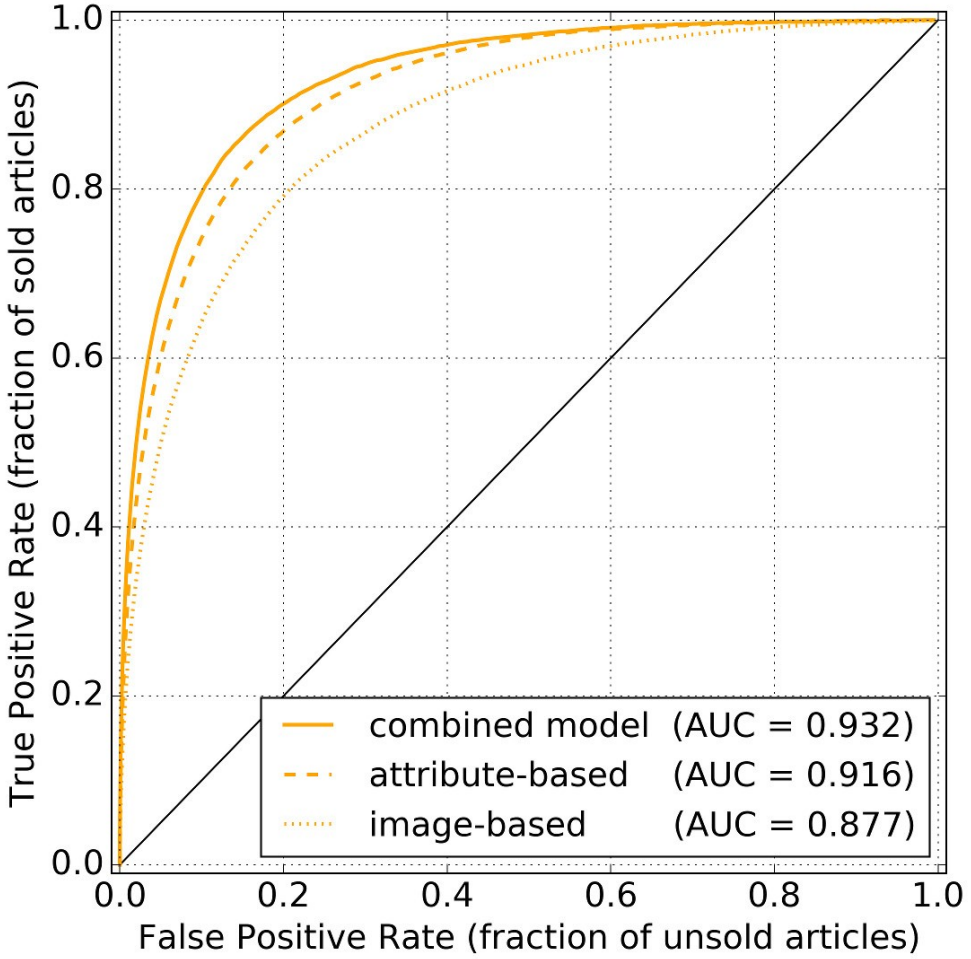
Distribution of Purchase Probabilities

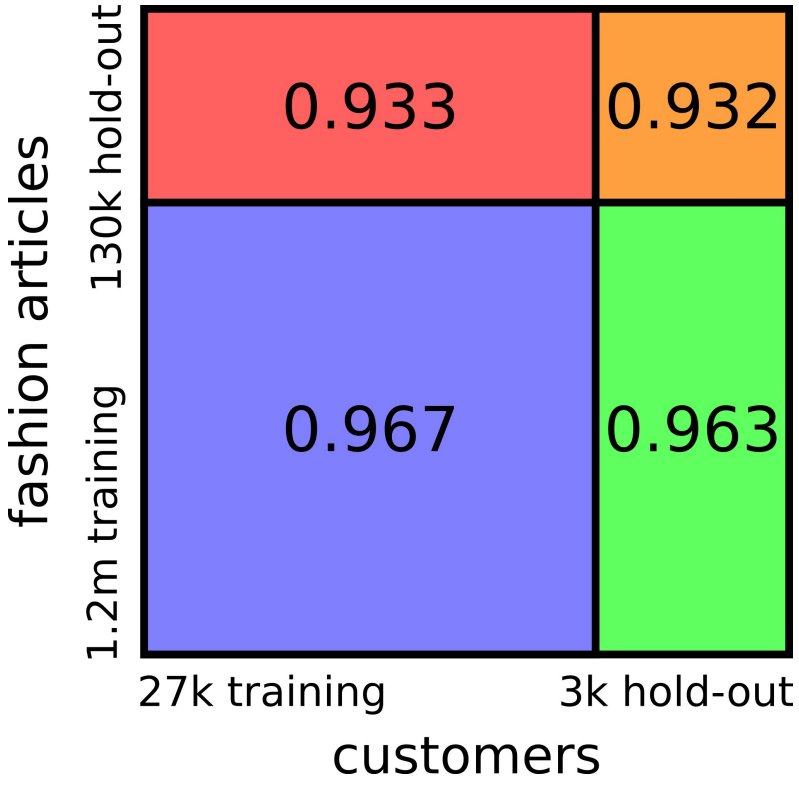
- Purchase probability for 200m customer-item pairs (validation set)
- Sorted into 200 bins
- Compare mean predicted probability to purchase rate in bin

- Forecast largely unbiased



ROC and AUC Analysis





Sample Recommendation: Hold-Out Customer

- Items bought in article hold-out set:

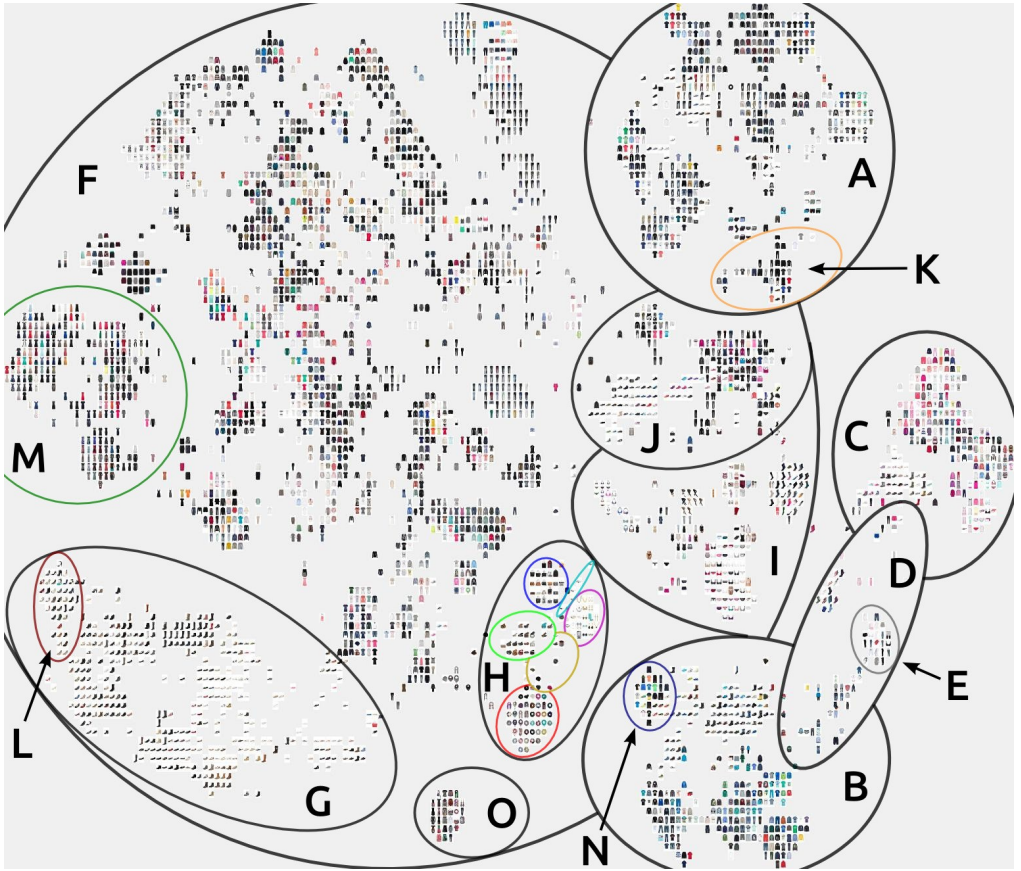
- Top recommendations in hold-out set:

- Purchases deemed least likely in hold-out set:

The Fashion Landscape: Nearest Neighbors

Sales data yields context

The Fashion Landscape: t-SNE Map



- A. **Men**
  - K. Mens' Sports
- B.-E. **Kids & Maternity**
  - B. Boys
    - N. Boys' Soccer Items
  - C. Girls
  - D. Babies & Toddlers
  - E. Maternity
- F. **Female**
  - G. Womens' Shoes
    - L. High Heels
  - H. Womens' Accessories
    - Scarves, Belts, Hats, Bags, Jewelry, Sunglasses
  - I. Lingerie, Hosiery, Swimwear
  - J. Women's Sports
  - M. Evening Dresses
  - O. Items by Desigual

Fashion DNA yields an hierarchical, orderly landscape of fashion items

Summary

- Our model integrates curated content, sales data
- Architecture: DNN & logistic matrix factorization

- Overcomes the cold start problem for new items
- Generalizes very well to hold-out customers
- Unbiased predictions for purchase probabilities

- Assigns vectors (Fashion DNA) to articles
- t-SNE map suggests hierarchical order in fashion space

Thanks to ...

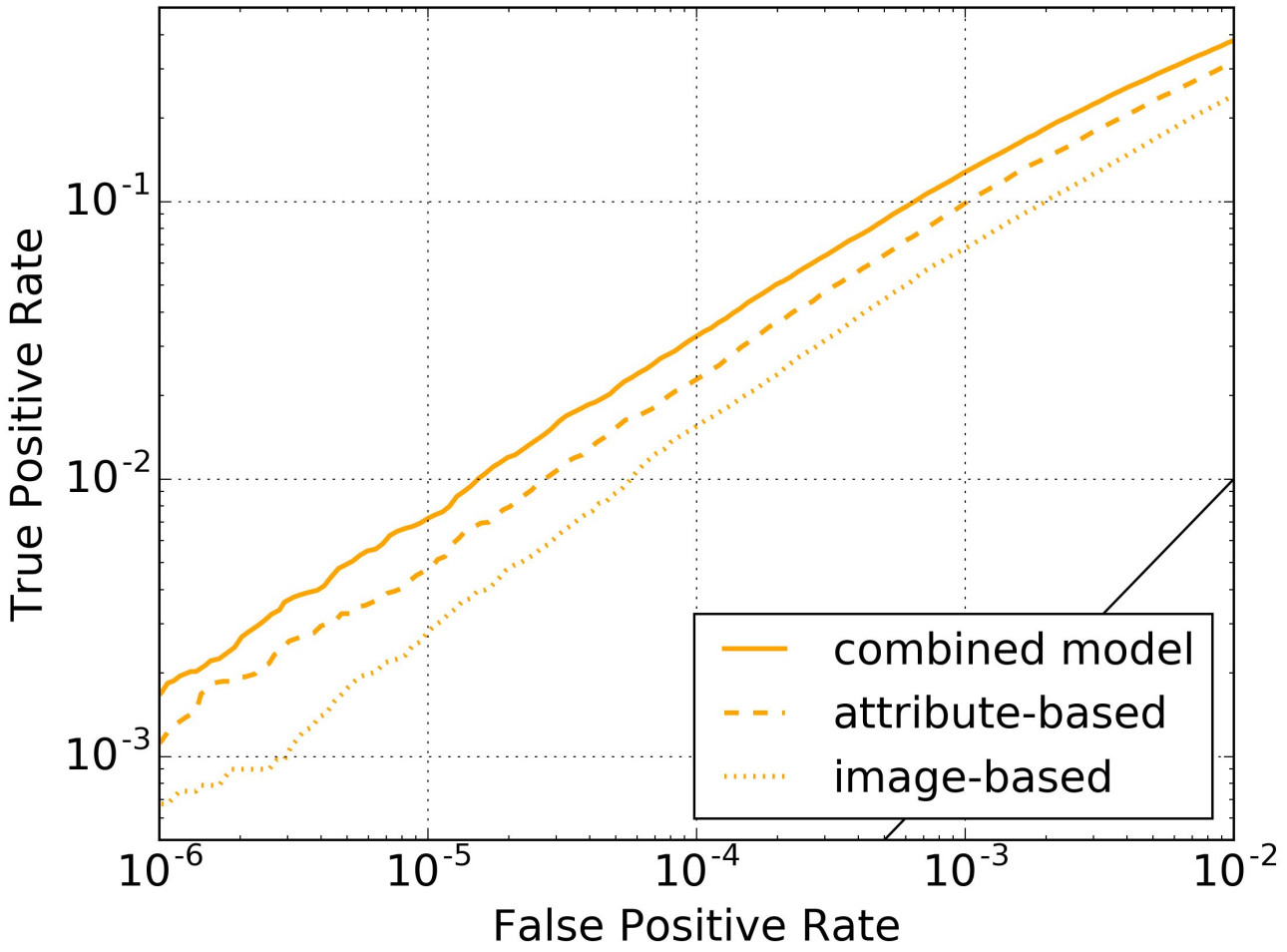
Sebastian Heinz **Roland Vollgraf**

Zalando Research, Berlin, Germany

Contact: {christian.bracher, sebastian.heinz, roland.vollgraf}@zalando.de

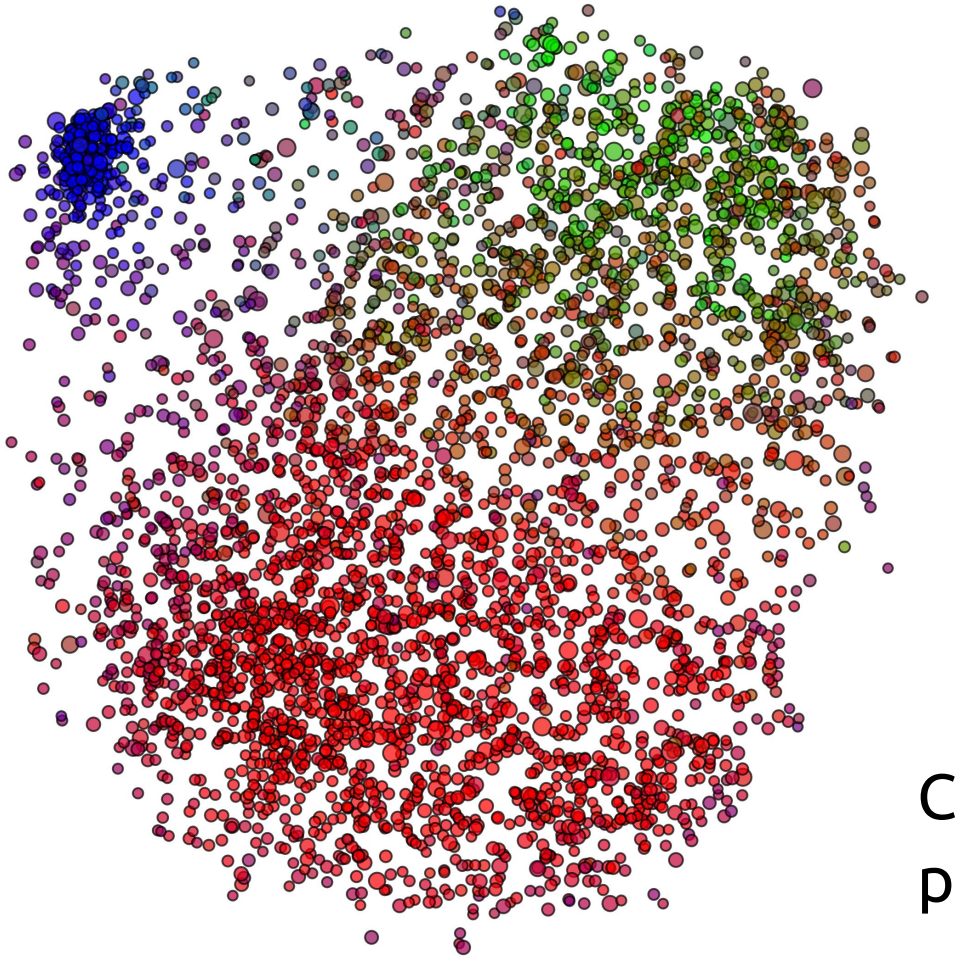

zalando
| RESEARCH |
THANK YOU!

ROC Analysis – Few Recommendation Limit



Nearest Neighbors: Model Dependence

The Customer Style Landscape: t-SNE Map



- 3000 frequent customers
- RGB color encoding by proportion of purchases:
  - womens' items (red)
  - mens' items (blue)
  - childrens' items (green)
- Size by number of sales

Customer styles beyond gender preference are idiosyncratic